Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 36, S. 237-242, Mayıs 2022
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 36, pp. 237-242, May 2022
Copyright © 2022 EJOSAT
Research Article

The Identification of Red-Meat Types using The Fine-Tuned Vision Transformer and MobileNet Models

Nagham Alhawas [1*], Zekeriya Tüfekci [2]

[1*] Çukurova University, Faculty of Engineering, Departmant of Computer Engineering, Adana, Turkey, (ORCID: 0000-0002-7407-1392), nagham_hs@hotmail.com
[2] Çukurova University, Faculty of Engineering, Departmant of Computer Engineering, Adana, Turkey, (ORCID: 0000-0001-7835-2741), ztufekci@cu.edu.tr

(1st International Conference on Engineering and Applied Natural Sciences ICEANS 2022, May 10-13, 2022)

(**DOI:** 10.31590/ejosat.1112892)

ATIF/REFERENCE: Alhawas, N. & Tüfekci, Z. (2022). The Identification of Red-Meat Types using The Fine-Tuned Vision Transformer and MobileNet Models. *European Journal of Science and Technology*, (36), 237-242.

Abstract

For reasons related to poverty or lack of quality control over food in some countries of the world, there is still food adulteration. Low-cost meats such as donkey or pork are marketed as lamb or beef. This is morally dangerous but may be more dangerous for some people who are allergic to certain types of meat or who have religious reservations. With the rapid development of artificial intelligence techniques, it is possible to build a model capable of differentiating between different types of meat. This study aims to build a model capable of differentiating between different types of red meat. It also aims to compare performance between the very state of art CNN in computer vision with the transformer architecture. For this goal, a limited dataset from an online repository was obtained. The dataset contains RGB images of beef, horse, and pork meats. The images were processed, and various data augmentation techniques were applied. Then vision transformer ViT and mobile net models with and without fine-tuning were built. To measure the models' behavior, several performance evaluation criteria were applied. The best testing accuracy is 97% achieved by the fine-tuned ViT model. This study showed the effectiveness of applying the transformer architecture and especially the fine-tuned ViT model in the areas of image classification even on a limited dataset.

Keywords: Computer vision, , fine-tuning, Vision Transformer, ViT, mobilenet.

İnce Ayarlı Görüntü Transformatörü ve MobileNet Modelleri Kullanılarak Kırmızı Et Türlerinin Belirlenmesi

Öz

Dünyanın bazı ülkelerinde yoksulluk veya gıda üzerinde kalite kontrolünün olmaması ile ilgili nedenlerden dolayı, hala gıda tağşişi var. Eşek veya domuz eti gibi düşük maliyetli etler kuzu veya sığır eti olarak pazarlanmaktadır. Bu ahlaki açıdan tehlikelidir, ancak belirli et türlerine alerjisi olan veya dini çekinceleri olan bazı kişiler için daha tehlikeli olabilir. Yapay zeka tekniklerinin hızla gelişmesiyle farklı et türleri arasında ayrım yapabilen bir model oluşturmak mümkün. Bu çalışma, farklı kırmızı et türleri arasında ayrım yapabilen bir model oluşturmayı amaçlamaktadır. Aynı zamanda, bilgisayarlı görü alanındaki en son teknoloji CNN ile transformatör mimarisi arasındaki performansı karşılaştırmayı da amaçlamaktadır. Bu amaç için, çevrimiçi bir depodan sınırlı bir veri seti elde edildi. Veri seti sığır, at ve domuz etlerinin RGB görüntülerini içermektedir. Görüntüler işlendi ve çeşitli veri büyütme teknikleri uygulandı. Daha sonra ince ayarlı ve ayarsız görüntü dönüştürücü ViT ve mobil ağ modelleri üretildi. Modellerin davranışını ölçmek için çeşitli performans değerlendirme kriterleri uygulandı. En iyi test doğruluğu, ince ayarlı ViT modeli tarafından elde edilen %97'dir. Bu çalışma, dönüştürücü mimarisinin ve özellikle ince ayarlı ViT modelinin sınırlı bir veri setinde bile görüntü sınıflandırma alanlarında uygulanmasının etkinliğini göstermiştir.

Anahtar Kelimeler: Bilgisayarla görme, ince ayar, Vision Transformer, ViT, mobilenet.

[*] Corresponding Author: nagham_hs@hotmail.com

1. Introduction

Around the world and throughout history, red meat has been a staple of human diets. The United States Department of Agriculture claims that all mammalian meats are classified as red meats because they contain higher myoglobin (Wikimedia Foundation, 2022). There are many types of red meat, such as beef, goat, pork, and others. With the diversity of red meat, there are concerns about identifying the types of red meat on the market. These phobias are frequently founded on religious beliefs, but they may simply be a result of a desire to avoid meat fraud. It's possible that what's on the label isn't what you're getting. So, the detection of food adulteration is important. Low-cost meats such as donkey or pig are marketed as lamb or beef in some countries, particularly those with insufficient meat quality control. This fraud is intended in foods. In another scenario, some Muslims living in countries where pork is sold may not be able to inquire or differentiate between the two types of meat. This distinction between varieties of red meat is significant to them since eating pig is prohibited by the Islamic religion, as well as several Christian and Jewish groups. Fortunately, by using various artificial intelligence approaches and leveraging technological advancements, it is now feasible to recognize the sort of red meat by merely snapping a picture of it at the store.

In (GC et al., 2021), researchers in United States used deep learning neural network for the classification of seven different beef cuts. As training, validation, and testing data, 1,113 beef cut colored photos gathered from google were utilized. The study used VGG16 and ResNet v2 to classify the beef cut images. The VGG16 shows better performance and scored 98.6% on 116 test images whereas ResNet achieved 95.7% testing accuracy.

In (Huang & Gu, 2022), the study used electronic nose data to extract number of features from a multichannel input matrix. The study presents a framework based on one dimensional CNN and random forest regressor for the quantitative identification of pork-addled beef.

In (Asmara et al., 2018), the RGB and GLCM properties of a backpropagation neural network utilized to identify beef and pig flesh. The system's classification accuracy is 89.57 %.

(Fitrianto & Sartono, n.d.) build CNN model to classify two types of red meat, the beef and pork. The study uses 2550 data as training data, and 450 for testing. The highest accuracy achieved by the model is 97.5%.

Adulteration of red meat is a worldwide issue that affects private sector integrity and threatens those with food allergy or religious convictions. In this study we aim to utilize recent sophisticated artificial intelligence techniques to help in the identification between three different 3000 images of meat types. Also, this study presents a comparison study between the performance of vision transformer ViT and mobile Net models in the image classification on small dataset.

2. Material and Method

The dataset and methodology used in this study are presented in the following sections.

2.1. Dataset

A public repository was used to obtain the imagery dataset (IQBAL AGISTANY, 2022). This dataset was released in February 2022 and has not yet been examined. It contains 365 RGB images for three red meat classes: horse, beef, and pork. Figure 1 shows samples of the images in the dataset.

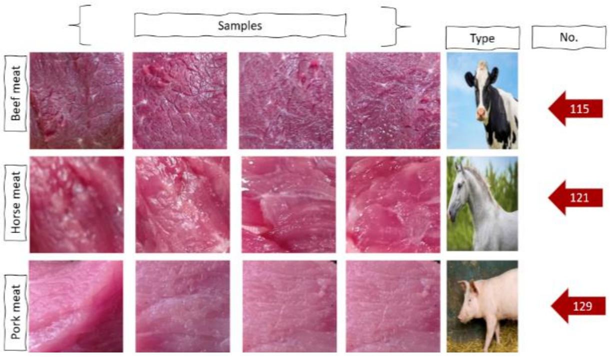


Figure 1 *Samples from the Imagery Dataset*

The dataset relatively small which present a real challenge for training deep learning model.

2.2. Methodology

Vision Transformer and MobileNet Models are used in the proposed methodology to develop a robust classifier that can distinguish between different varieties of red meat.

2.2.1 Vision Transformer ViT

The ViT model for vision task was first introduced in 2021 as a research paper at the ICLR 2021 conference (Dosovitskiy et al., 2020). The model appeared in 2022 as a good alternative and competitive to convolutional neural networks, which consider the state-of-art in computer vision (Zhou et al., 2021). The use of the ViT model in image classification tasks is very successful in the last two years and according to ImageNet competition in 2022, Vit is ranked as the best model for image classification tasks with 90.94% model accuracy and 1843M parameter and it takes 2.5k TPUv3-days. Figure 2 shows these results for all models. For this success in vision tasks, we decide to use the fine-tuned ViT which was trained on the ImageNet dataset.

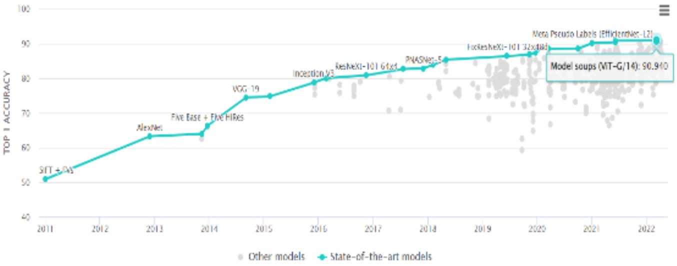


Figure 2 *Leader dashboard in Image Classification on ImageNet* (paperswithcode, 2022)

Without employing convolution layers, the ViT method utilizes the Transformer architecture for sequences of image patches. First the images in the dataset should be split into fixed sizes (patches). It creates lower-dimensional linear embeddings from these flattened image patches after flattening them. It uses a state-of-the-art transformer encoder to input the sequence. The overall architecture for the ViT model is shown in Figure 3.

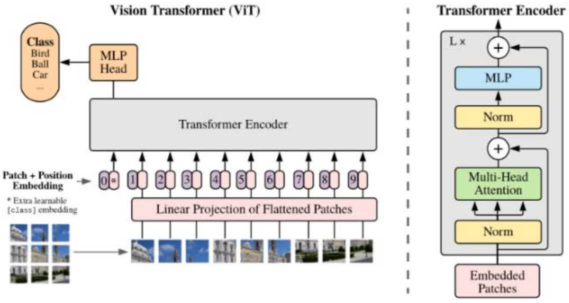
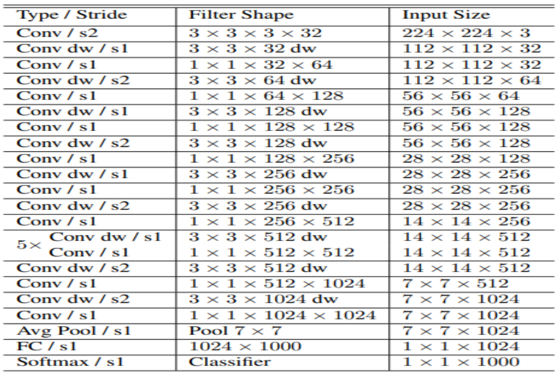


Figure 3 *The Overall architecture of ViT model (Dosovitskiy et al., 2020)*

2.2.2 MobileNet

In CVPR 2017, a group of Google engineers released MobileNet which is known as a lightweight model. Apart from the first layer, which is a full convolutional layer, it is also known as Depth wise Separable Convolution Network since it serves as the network's main building block. The network executes a single convolution on each RGB input channel independently, using the mean of depth-wise and pointwise convolutions. Unlike previous models that filter, and merge data based on convolutional kernels in one step to build a new representation, the MobileNet filters and combines data in two distinct stages(Howard et al., 2017). The summary of the mobileNet layers architecture is shown in Figure 4.



Type / Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw / s1	3 × 3 × 32 dw	112 × 112 × 32
Conv / s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw / s2	3 × 3 × 64 dw	112 × 112 × 64
Conv / s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw / s1	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw / s2	3 × 3 × 128 dw	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw / s1	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw / s2	3 × 3 × 256 dw	28 × 28 × 256
Conv / s1	1 × 1 × 256 × 512	14 × 14 × 256
5× Conv dw / s1	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw / s2	3 × 3 × 512 dw	14 × 14 × 512
Conv / s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw / s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv / s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool / s1	Pool 7 × 7	7 × 7 × 1024
FC / s1	1024 × 1000	1 × 1 × 1024
Softmax / s1	Classifier	1 × 1 × 1000

Figure 4 *The Summary of MobileNet model (Howard et al., 2017)*

The deep learning model's performance may be judged better with a larger dataset. Because we have a limited dataset, the transfer learning and fine-tuning strategy is used, which is the most popular deep learning trick when dealing with a limited dataset. It makes use of prior knowledge while freezing some layers and training the last layers with a low learning rate to adjust the model to the new role.

2.3. Experiments

As discussed in the methodology section, ViT model and MobileNet models were used. Even though the dataset is considered limited, the experiments on the two models were applied with and without fine-tuning to see the differences in performance. The software and hardware configuration that were used to complete the project's experiments is shown in **Hata! Başvuru kaynağı bulunamadı.**.

Table 1 Project's software and hardware parameters

Hardware / software	**Parameter**
Operating System	*Windows 11 pro× 64*
CPU	*11th Gen Intel® Core ™*
GPU	*NIVIDIA GeForce RTX*
Programming language	*Python*
IDE	*Jupyter*
Deep learning library	*TensorFlow*

The dataset is split into training, validation, and testing. The training set contains 335 images and 30 images of testing. The validation data is separated from training data with validation_split () fraction equal to 0.1. Figure 5 shows the total images for training per class.

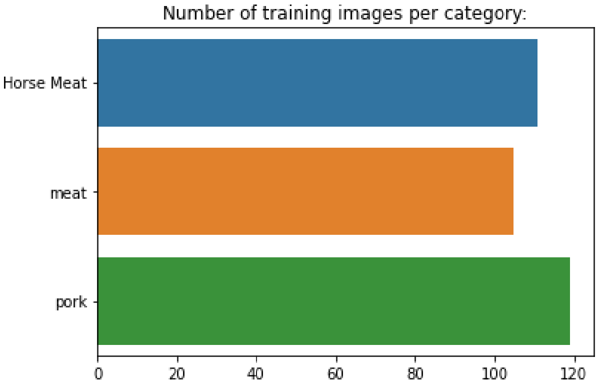


Figure 5 *The number of the training images per category*

Data augmentation techniques such as pixels normalization, zooming, and rotating were used to increase the number of images to avoid the model's overfitting(Kaur et al., 2021). The training details and evaluation matrices are done as follows:

2.3.1. The Models' Training Details

2.3.1.1. Building and training ViT Model

Even though we are interested in the fine-tuned ViT model, a ViT model without fine-tuning was built from scratch as well to compare the performance of both. For the simple ViT model, the hyperparameter of the model were all configured precisely. The images were resized to 72×72 and the projection dimension is 64. The transformer layer size was configured as 2× projection dimensions. After that, the multilayer perceptron was implemented with two layers with Gaussian Error Linear Unit (GELU). ViT model uses the transformer architecture with self-attention layer to represent the input image as sequence of patches to perform on it. Patch layer is implemented with 6 X 6 patch size, 144 patches per image, and 108 elements per patches. The result of visualizing one image under this consideration is shown in Figure 6.

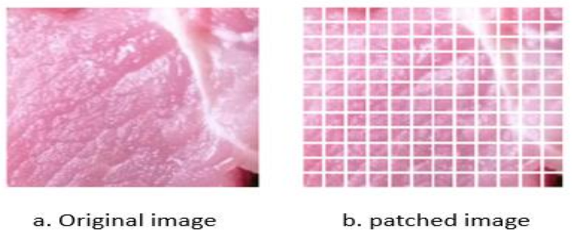


Figure 6 *Sample result of using the patch layer*

Then the patch encoder layer was implemented, which will linearly transform a patch into a vector equal to the size of projection dimension that was sited before and add a learnable position embedding to it. Finally, the ViT model were built with multiple transformer blocks with SoftMax to produce the final probabilities output the model was trained with 100 epochs.

For the fine-tuned model, the vit_b32 model from vit_keras was loaded without including the top with SoftMax as activation function and the image size set to 224 ×224 (Andreas Steiner, 2022). One of the fundamental components of transformers is attention, especially self-attention. It's a computational primitive that helps a network understand the hierarchies and alignments found in input data by quantifying paired entity interactions(Gaudenz Boesch, 2022). Figure 7 shows the result of visualizing the attention maps for two test sample images.

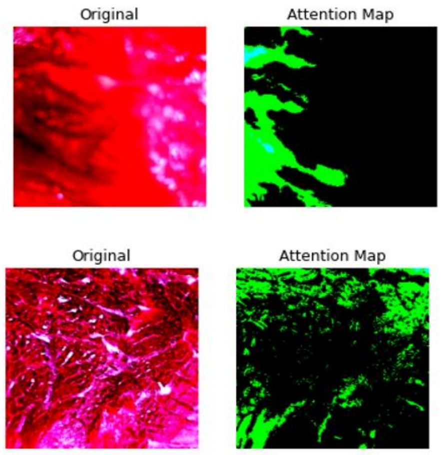
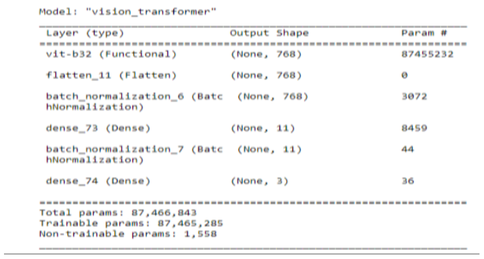


Figure 7 *Visualization of attention map for test images*

Then the fine tuned ViT model is built with the retrained weights on imagenet data set. The model compiled with learning rate is 1 x 10-4, Rectified Adam as an optimizer, and number of epochs equal 100. Figure 8 shows the full summary of the fine-tuned model.



```
Model: "vision_transformer"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vit-b32 (Functional)        (None, 768)               87455232

 flatten_11 (Flatten)        (None, 768)               0

 batch_normalization_6 (Batc  (None, 768)              3072
 hNormalization)

 dense_73 (Dense)            (None, 11)                8459

 batch_normalization_7 (Batc  (None, 11)               44
 hNormalization)

 dense_74 (Dense)            (None, 3)                 36

=================================================================
Total params: 87,466,843
Trainable params: 87,465,285
Non-trainable params: 1,558
_________________________________________________________________
```

Figure 8 *The fine-tuned ViT model summary*

2.3.1.2. Building and Training Mobile Net Model

The image tensors were processed with the mobile pre-processing function to be fed to the model. The pretrained mobile net model is loaded with image net weights and the last layers were cut off and re-trained w.ith small learning rate with Adam as an optimizer. The total number of parameters is 3,231,939 and only 1,860,099 trainable parameters. This is the fine-tuned version of the mobilenet model, however, the model without fine tuning were used which means all the 3,231,939 parameters were set as trainable and without any pretrained weights. This is done to compare the performance of each case.

2.3.2. Evaluation Matrices

Different performance assessment matrices are utilized to assess each model's performance. Model accuracy is defined as the number of classifications a model correctly predicts divided by the total number of predictions. The precision is a percentage of relevant instances (true positives) among all the examples expected to belong to a particular class. The recall is defined as the percentage of instances predicted to belong to a class divided by the total number of examples that belong to that class. The equations used to measure the accuracy, precision, recall, and f1 equations are illustrated in 2,3,4,5 Eq. The testing data's confusion matrix is plotted for the two models. The Confusion Matrix is a statistic that describes the prediction performance of a machine learning model. Knowing the difference between true positives, false positives, true negatives, and false negatives is very helpful to understand the model's performance. The analogy of the used confusion matrix is shown in Figure 9. The time taken for each model to complete the training phase is measured to compare the performance of each one. As a final performance measurement, the model accuracy and loss during the training phase were plotted.

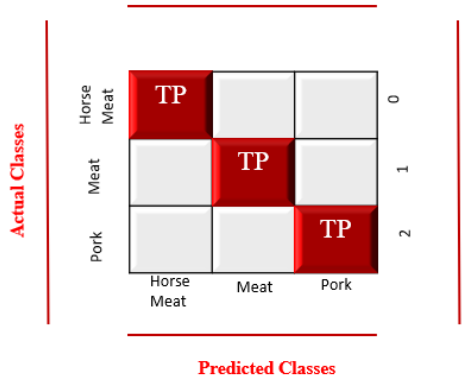


Figure 9 *Confusion Matrix for the three classes*

$$\text{Accuracy} = \frac{\text{correct predictions}}{\text{all predictions}} \text{ (Eq. 2)}$$

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad \text{(Eq. 3)}$$

$$Recall = \frac{True\ positives}{true\ positives + false\ positives} \quad \text{(Eq.4)}$$

$$F1 = 2 \times \frac{Percision \times recall}{precision + recall} \quad \text{(Eq. 5)}$$

3. Results

Using advanced techniques, two main experiments were used to classify the three red meat classes. The same software and hardware setups were utilized in all experiments, as shown in **Hata! Başvuru kaynağı bulunamadı.**. Distinguishing beef, horse, and pork cuts meat can be a delicate process even using the naked eye. Therefore, the challenge to obtain the most accurate distinction classifier between the three types is the criterion for the success of this study. As we mentioned in the previous section, the ViT model was built from scratch, and on the side, the Fine-Tuned ViT model was used and built. As for the ViT model that was built from scratch, it took 14:13 minutes for to train the model with 100 epochs. This model achieved 97.67% training accuracy, 91.18% for validation, and 83.3% testing accuracy. While the fine-tuned ViT model took about 2 hours to train the model with 100 epochs. The model gives extremely better performance than the ViT model without fine-tuning. It scored 100% accuracy for training and validation. For testing the model scores 97%. It was noticeable that the performance of the fine-tuned ViT model did not improve after the 68 epochs, and the Accuracy remained conservative and stable at 100%. With an early stop, the training could have been stopped, but we wanted to compare the performance of all models with the 100 epochs. The confusion matrix for the fine-tuned ViT model on testing image samples of the three classes is shown in Figure 10. When looking at these results, only one of the other ten images of beef was misclassified as pork, which is an acceptable amount of error.

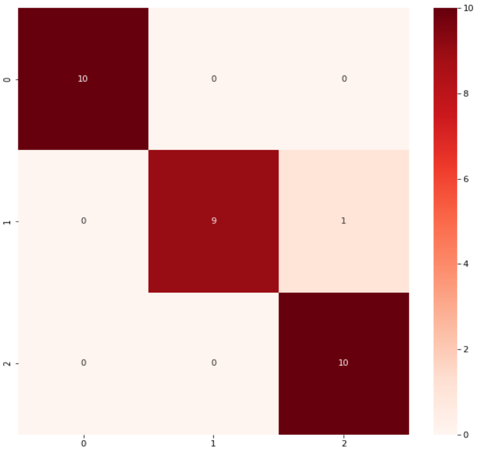
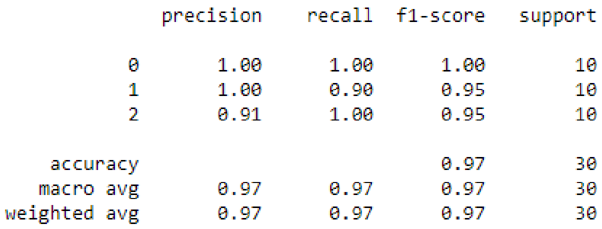


***Figure 10** The confusion matrix of the fine-tuned ViT model*

The full classification report for the fine-tuned ViT model is illustrated in Figure 11.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        10
           1       1.00      0.90      0.95        10
           2       0.91      1.00      0.95        10

    accuracy                           0.97        30
   macro avg       0.97      0.97      0.97        30
weighted avg       0.97      0.97      0.97        30
```

***Figure 11**Classification results for the fine-tuned ViT model*

On the other hand, the result of mobilenet model without fine tuning was not good enough to rely on the model. The model took 25:52 minute for training stage and scored 99% for training accuracy and 100% for validation. However, the model achieved fail in testing and give 33% accuracy which indicate overfitting. While the fine-tuned mobilenet gives better performance it took 11:15 minute to achieve 100% for training and validation accuracy. Also, it's clear that the accuracy didn't change after the 90th epoch and remain 100% for training and validation as shown in Figure 12.

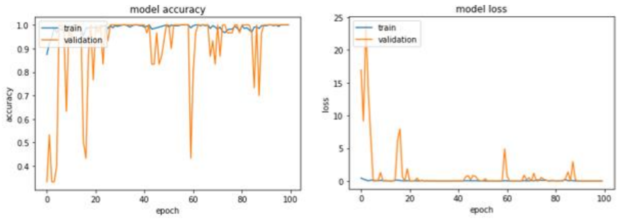


***Figure 12** The Fine-tuned mobilenet model training accuracy and loss*

The model scored 83% for testing accuracy. The confusion matrix of the model on the tree classes for the testing samples is shown in Figure 13.

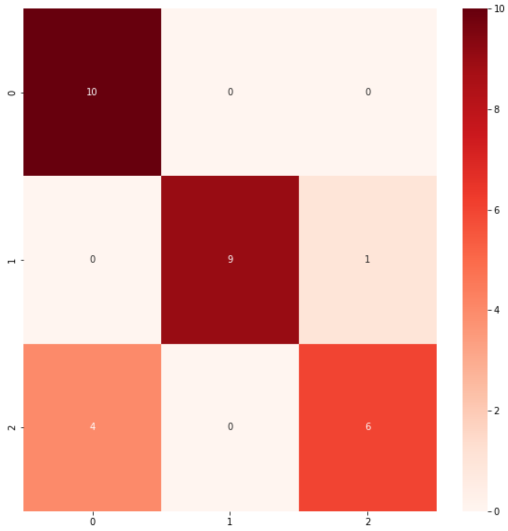
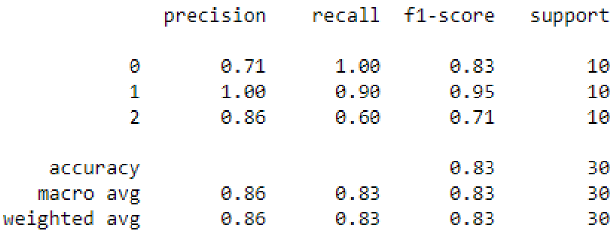


***Figure 13** Confusion matrix of fine-tuned mobilenet*

The complete classification results of the fine-tuned mobilenet model are shown in Figure 14.



```
              precision    recall  f1-score   support

           0       0.71      1.00      0.83        10
           1       1.00      0.90      0.95        10
           2       0.86      0.60      0.71        10

    accuracy                           0.83        30
   macro avg       0.86      0.83      0.83        30
weighted avg       0.86      0.83      0.83        30
```

***Figure 14** classification results for the fine-tuned mobilenet model*

4. Discussion

The results of the conducted experiments to classify 3 different types of red meat types showed that the study achieved its goals. Two fine-tuned classifiers ViT and mobilenet models were built. These experiments were preceded by the build of the same two classifiers without fine-tuning. However, the results showed that the use of both classifiers without fine-tuning led to overfitting and so, the models fail to learn on the test data. This is not surprising for us since we have a limited dataset. Therefore, in cooperating fine-tuning with ViT and mobilenet models gives better performance. The fine-tuned ViT model achieved better performance than fine-tuned mobilenet. It scored 97% for testing whereas the fine-tuned mobilenet achieved 83.3%. No doubt fine-tuned ViT model took longer time for training compared to the fine-tuned mobilenet model. There is a huge difference between the number of parameters in both models. The fine-tuned ViT model has 87,465,285 trainable parameters whereas the fine-tuned mobilenet has 1,860,099 trainable parameters and this difference to an expected difference in the time required to train each model. However, it was enough for half of 100 epochs to reach the same result for ViT model, but this is not the case for the fine-tuned mobilenet model. The results comparison between the two fine-tuned models with 100 epochs is shown in Table 2.

Table 2 The result of fine-tuned ViT and mobilenet models

Model	**Training accuracy**	**validation accuracy**	**Testing accuracy**	**time**
Fine-tuned Vit	100%	100%	97%	2:50 h
Fine-tuned mobilenet	100%	100%	83%	11:15m

The Experiences have shown that vision transformer architecture, in the fine-tuned model is indeed a model that competes with CNN in the field of computer vision and could perform even better.

5. Conclusion

Some stores located in some poor countries or those suffering from wars or others, resort to selling pork as beef for the cheaper price of the former compared to the latter. In some religions, there are some types of red meat that are forbidden to their adherents. In one way or another, determining the type of meat displayed in stores is considered a matter that contributes to food security and enhances the credibility of sellers. After the great development in computer vision techniques and artificial intelligence, it became possible to create a program that enables users to determine the types of meat displayed by taking a picture of the meat only. In this study, we constructed fine-tuned ViT model to differentiate between beef, horse, and pork. The classifier achieved 97% testing accuracy despite the small and limited dataset. This model can be used for study reasons or to make applications to differentiate between types of meat for religious, health, or ethical reasons related to consumer food security.

References

Andreas Steiner. (2022). *Vision Transformer and MLP-Mixer Architectures*. Https://Github.Com/Google-Research/Vision_transformer.

Asmara, R. A., Romario, R., Batubulan, K. S., Rohadi, E., Siradjuddin, I., Ronilaya, F., Ariyanto, R., Rahmad, C., & Rahutomo, F. (2018). Classification of pork and beef meat images using extraction of color and texture feature by Grey Level Co-Occurrence Matrix method. *IOP Conference Series: Materials Science and Engineering*, *434*(1). https://doi.org/10.1088/1757-899X/434/1/012072

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. http://arxiv.org/abs/2010.11929

Fitrianto, A., & Sartono, B. (n.d.). *International journal of science, engineering, and information technology Image Classification of Beef and Pork Using Convolutional Neural Network in Keras Framework*. https://journal.trunojoyo.ac.id/ijseit

Gaudenz Boesch. (2022). *Vision Transformers (ViT) in Image Recognition – 2022 Guide*. Https://Viso.Ai/Deep-Learning/Vision-Transformer-Vit/.

GC, S., Saidul Md, B., Zhang, Y., Reed, D., Ahsan, M., Berg, E., & Sun, X. (2021). Using Deep Learning Neural Network in Artificial Intelligence Technology to Classify Beef Cuts. *Frontiers in Sensors*, *2*. https://doi.org/10.3389/fsens.2021.654357

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. http://arxiv.org/abs/1704.04861

Huang, C., & Gu, Y. (2022). A Machine Learning Method for the Quantitative Detection of Adulterated Meat Using a MOS-Based E-Nose. *Foods*, *11*(4). https://doi.org/10.3390/foods11040602

IQBAL AGISTANY. (2022, February). *Pork, Meat, and Horse Meat Dataset*. Https://Www.Kaggle.Com/Datasets/Iqbalagistany/Pork-Meat-and-Horse-Meat-Dataset.

Kaur, P., Khehra, B. S., & Mavi, Er. B. S. (2021). Data Augmentation for Object Detection: A Review. *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, 537–543. https://doi.org/10.1109/MWSCAS47672.2021.9531849

paperswithcode. (2022, March 2). *Image Classification on ImageNet*. Https://Paperswithcode.Com/Sota/Image-Classification-on-Imagenet?P=centroid-Transformers-Learning-to-Abstract.

Wikimedia Foundation. (2022, March 3). *Red meat*. Https://En.Wikipedia.Org/Wiki/Red_meat.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., & Feng, J. (2021). *DeepViT: Towards Deeper Vision Transformer*. http://arxiv.org/abs/2103.11886